20269 – Economics of European Integration – 2021/22

Take Home

MAXIMUM ALLOWED SPACE = 25 PAGES INCLUDING TABLES[1]

The Stata dataset "*EEI_TH_2022.dta*" can be downloaded from the course Blackboard. It contains data on more than 80,000 European firms operating in 3 countries (Spain, France and Italy) and in 2 industries (*Manufacture of textiles* and *Manufacture of motor vehicles, trailers and semi-trailers*, coded according to the official NACE rev. 2 nomenclature as industry 13 and 29 respectively), over the period 2000-2017.

The variables are included in the dataset in levels and are the following:

- *year* and *country*
- *nace*: industry (NACE rev. 2 classification, codes 13 and 29) in which the firms operate
- *nuts2*: region (NUTS 2) in which the firms operate
- *id_n*: a code identifying each firm
- *sizeclass*: the size class of firms (depending on number of employees)
- *L*: the number of workers
- *K*: the capital input (in thousands Eur) and its deflated value (*real_K*)
- *M*: the raw materials input (in thousands Eur) and its deflated value (*real_M*)
- *W*: the total labor costs (in thousands Eur) paid by the firm
- *sales*: the revenues of the firm (in thousands Eur) and its deflated value (*real_sales*)
- *real_VA*: a proxy measuring the deflated value added of the firm (where value added = revenues – materials)

Using this dataset, you are requested to solve the following problems:

Problem I

a. **Focus on Italian firms only**. Starting from balance-sheet data, provide some descriptive statistics in 2008 for the firms in the sample (e.g. n. of firms, average capital, revenues, number of employees, and other variables that you may consider as relevant) by industry. Comment briefly.

b. Compare the descriptive statistics that you have analysed for 2008 to the same figures in 2017 for the same country. What changes? Comment and give an interpretation.

[1] Please send your work in PDF format to Lorenzo Cavaglià (lorenzo.cavaglia@unibocconi.it) before the deadline. Please also attach your STATA .do and final .dta files. Please send all your outputs in a single email. For any questions or doubts, please consider Lorenzo Cavaglià as your main reference for this Take Home.

Problem II

a. **Consider now all the three countries**. Estimate for the two industries available in NACE Rev. 2 2-digit format the production function coefficients, by using standard OLS, the Wooldridge (WRDG) and the Levinsohn & Petrin (LP) procedure.
   How do you treat the fact that data come from different countries in different years in the productivity estimation?
b. Present a Table (like the one below), where you compare the coefficients obtained in the estimation outputs, indicating their significance levels (*, ** or *** for 10, 5 and 1 per cent). Is there any bias of the labour coefficients? What is the reason for that?

		Nace-13	Nace-29
Lev-Pet	ln(labor)		
	ln(capital)		
WRDG	ln(labor)		
	ln(capital)		
OLS	ln(labor)		
	ln(capital)		
	Bias in labour coefficient		
	N. of observations		

Problem III

a. Would there be any difference in estimating the production function using revenues rather than added values in LP, WRDG or OLS? Why is it so? Discuss the issue theoretically, considering the assumptions behind the Cobb-Douglas production function.

Problem IV

a. Comment on the presence of "extreme" values in both industries. Clear the TFP estimates from these extreme values (1st and 99th percentiles) and save a "cleaned sample". **From now on, focus on this sample.** Plot the kdensity of the TFP distribution and the kdensity of the logarithmic transformation of TFP in each industry. What do you notice? Are there any differences if you rely on the LP or WRDG procedure? Comment.
b. Plot the TFP distribution for each country. Are there any differences if you rely on the LP or WRDG procedure? Compare and comment.
c. Focus now on the TFP distributions of industry 29 in France and Italy. Do you find changes in these two TFP distributions in 2001 vs 2008? Did you expect these results? Compare the results obtained with WRDG and LP procedure and comment.
d. Look at changes in skewness in the same time window (again, focus on industry 29 only in these two countries). What happens? Relate this result to what you have found at point c.

e. Do you find the shifts to be homogenous throughout the distribution? Once you have defined a specific parametrical distribution for the TFP, is there a way through which you can statistically measure the changes in the TFP distribution in each industry over time (2001 vs 2008)?

****

For the second part of this Take Home, you first need to recall how to compute the China shock at the regional level. To this purpose, you can refer to the formula by Colantone and Stanig (*American Journal of Political Science*, 2018):

$$ChinaShock_{crt} = \sum_k \frac{L_{rk(pre-sample)}}{L_{r(pre-sample)}} * \frac{\Delta IMPChina_{ckt}}{L_{ck(pre-sample)}} \qquad (1)$$

link to micrometrics PS2

where *c* indexes countries, *r* regions, *k* industries in the manufacturing sector, and *t* years.

You now need four more datasets (all downloadable from Blackboard).

The Stata dataset "*Employment_Shares_Take_Home.dta*" contains employment data for each NACE rev.1.1 industry in Italian, French and Spanish regions. As per the equation above, the employment variables refer to the earliest pre-sample year in which they could be observed. As such, they display the same values for each region and industry over time. The database is nevertheless structured as a panel (from 1988 to 2007) to facilitate the merge with the other two datasets presented below. You may use this dataset to compute weights for import deltas, i.e., the first term and the denominator of the second term of equation (1).

The variables are included in the dataset in levels and are the following:

- *year* and *country* of each observation
- *nuts_code*: the region code; *nuts_name*: the region name
- *nace*: the NACE rev.1.1 industry
- *L_rk*: number of employees by region-industry pre-sample
- *L_r*: total number of employees by region pre-sample
- *L_ck*: total number of employees by country-industry pre-sample

The Stata dataset "*Imports_China_Take_Home.dta*" contains data on imports from China to Italy, France and Spain for each NACE rev.1.1 industry in each year, from 1988 to 2007. You may use this dataset to compute import deltas, i.e., the numerator of the second term of equation (1).

The variables are included in the dataset in levels and are the following:

- *year* and *country* of each observation
- *nace*: the NACE rev.1.1 industry
- *real_imports_china*: deflated imports from China

The Stata dataset "*Imports_US_China_Take_Home.dta*" contains data on imports from China to the US for each NACE rev.1.1 industry in each year, from 1989 to 2006. You may use this dataset to construct an instrumental variable, which is needed in the remaining part of the Take Home.

The variables are included in the dataset in levels and are the following:

- *year* of each observation
- *nace*: the NACE rev.1.1 industry
- *real_USimports_china*: deflated imports from China to the US

Finally, the Stata dataset "*EEI_TH_5d_2022_V2.dta*" contains data on European industries at the regional (NUTS-2) level in 3 countries (Spain, France and Italy) over the period 2000-2017. You may use this dataset in the last part of Problem V. The included variables are:

- *year and country* of each observation
- *nuts_code*: the region code
- *nace*: the NACE rev.1.1 industry
- *tfp*: mean TFP at the NUTS-2 and industry level (obtained from firm-level data)
- *mean_uwage*: mean wage at the NUTS-2 and industry level (obtained from firm-level data)
- *lnpop*: population of the region in log
- *control_gdp*: gdp growth (%) of the region
- *share_tert_educ:* share of population with tertiary education (%) of the region

Problem V

a. Merge the first three datasets together. Compute the China shock for each region, in each year for which it is possible, according to equation (1). Use a lag of 5 years to compute the import deltas (i.e., growth in imports between *t-6* and *t-1*).
   Repeat the same procedure with US imports, i.e., substituting $\Delta IMPChina_{ckt}$ with $\Delta IMPChinaUSA_{kt}$, following the identification strategy by Colantone and Stanig (*AJPS*, 2018).

b. Collapse the dataset by region to obtain the average 5-year China shock over the sample period. This will be the average of all available years' shocks (for reference, see Colantone and Stanig, *American Political Science Review*, 2018). You should now have a dataset with cross-sectional data.

c. Using the cross-sectional data, produce a map visualizing the China shock for each region, i.e., with darker shades reflecting stronger shocks. Going back to the "*Employment_Shares_Take_Home.dta*", do the same with respect to the overall pre-sample share of employment in the manufacturing sector. Do you notice any similarities between the two maps? What were your expectations? Comment.

Problem VI

Use the dataset "*EEI_TH_5d_2022_V2.dta*" to construct, for each NUTS-2 and industry level, an average of tfp and wages during the post-crisis years (2014-2017). These will be your dependent variables. Now merge the data you have obtained with data on the China shock (region-specific average).

a. Regress (simple OLS) the post-crisis average of tfp against the region-level China shock previously constructed. Use population, education and gdp set at the beginning of the period in which your dependent variable is measured (2014) as controls. Comment on the estimated coefficient on the China shock, and discuss possible endogeneity issues.
b. To deal with endogeneity issues, use the instrumental variable you have built before, based on changes in Chinese imports in the USA, and run again the regression as in a). Do you see any change in the coefficient?
c. Now, regress (both OLS and IV) the post-crisis average of wage against the region-level China shock previously constructed. Use population, education and gdp set at the beginning of the period in which your dependent variable is measured (2014) as controls. Comment on the estimated coefficient on the China shock, and discuss possible endogeneity issues.
d. Lastly, run again the regression as in c), but now also add the average of tfp during the post-crisis years (the dependent variable of regressions a) and b)) as a control. Do you see any change in the coefficient of the China shock? Comment.

****

For the final part of this Take Home, you are asked to focus **only on Italian regions**.

First, please go to the European Social Survey official website and download the free-access Round 8 data. From this wave, keep only Italian respondents. Also, keep the following variables:

- Survey weights (post-stratification weight including design weight)
- Gender
- Age
- Highest level of education
- Region (NUTS 2)
- Party voted for in last national elections

Having constructed your dataset from ESS, solve the following problems:

Problem VII

a. Merge the data you have obtained from ESS with data on the China shock (region-specific average).

b. Create a dummy equal to one if the respondent has voted for a radical-right party in the last elections. That is, either Lega Nord or Fratelli d'Italia. Regress (simple OLS) this dummy against the region-level China shock previously constructed, controlling for gender, age, and dummies for levels of education. Cluster the standard errors by region. Be sure to use survey weights in the regression. Comment on the estimated coefficient on the China shock, and discuss possible endogeneity issues.
c. To correct for endogeneity issues, use the instrumental variable you have built before, based on changes in Chinese imports in the USA. Discuss the rationale for using this instrumental variable. What happens when you instrument the China shock in the previous regression? Comment both on first-stage and on second-stage results.
d. Do you notice any bias in the OLS estimates with respect to the IV ones? Comment.